第48卷 第1期
2025年1月

干旱区地理
ARID LAND GEOGRAPHY

Vol. 48 No. 1
Jan. 2025

内蒙古冰雹特征及基于机器学习的冰雹识别方法研究

辛 悦[1]， 苏立娟[1]， 郑旭程[1]， 李 慧[1]， 衣娜娜[1]， 靳雨晨[2]

（1. 内蒙古自治区人工影响天气中心，内蒙古 呼和浩特 010051； 2. 内蒙古自治区气象科学研究所，内蒙古 呼和浩特 010051）

摘 要：利用1959—2021年内蒙古人工观测冰雹记录，分析冰雹分布的时空特征，并基于机器学习算法构建了冰雹识别方法。结果表明：（1）时间分布上，冰雹事件出现的站数和站日数均呈现下降趋势；空间分布上，冰雹多集中在阴山山脉和大兴安岭一带，冰雹多发区沿山脉伸展分布。（2）冰雹发生具有明显的季节变化和日变化特征，每年5—9月是冰雹频发月份，占全年雹日的91.79%，雹日中12:00—19:00是冰雹的多发时段。（3）利用随机森林、LightGBM、*K*近邻和决策树4种机器学习算法，通过数据预处理、预报因子选择、模型训练、模型调优等步骤，对内蒙古冰雹天气过程进行建模与评估。评估结果表明，采用机器学习方法可以有效地识别冰雹天气过程，各模型的TS评分均达到0.83以上，命中率达到92%以上，随机森林算法在测试集上识别效果最优。研究结果可为内蒙古冰雹预报预警和人工防雹工作提供参考。

关 键 词：冰雹站日数；时空特征；机器学习；冰雹识别

文章编号：1000-6060(2025)01-0011-09(0011~0019)

冰雹作为一种短时强灾害天气，对农业生产、交通运输、建筑设施甚至生命财产安全等造成严重影响[1-3]。由于冰雹灾害性天气具有出现范围小、时效较短、来势凶猛、突发性强等特点，其预报预警一直是气象业务中的难点。为更好地理解冰雹的发生规律对其实现提前预报预警，国内学者在冰雹时空演变的气候态特征、识别指标方法研究、冰雹形成机理等方面开展了大量研究工作。

张芳华等[4]利用1971—2000年614个台站的冰雹观测资料，分析了冰雹日数的分布特征及其时间变化规律，发现我国冰雹频发区主要为青藏高原、华北东北部和东北地区，近30 a我国冰雹日数有显著减少的趋势，冰雹发生有明显的季节变化和日变化特征，夏季和冬季分别为冰雹站日数最多和最少的季节，冰雹发生时段主要集中在午后至傍晚。汤兴芝等[5]对2010—2020年我国冰雹灾害事件时空分布特征进行研究，得到了与张芳华等[4]相似的结论。尉英华等[6]统计分析了天津地区近11 a冰雹时空特征，对比分析了不同月份、不同天气型及不同大小冰雹的环境条件差异，给出了具有表征意义的环境参量特征值作为冰雹预报指标。钟敏等[7]利用NCEP GFS资料统计了湖北省25次冰雹天气过程的天气形势分类和环境场物理量，利用二分法和连续概率方法，建立了不同天气形势下的冰雹发生概率预报模型，并实现业务运行。刘新伟等[8-9]基于C波段雷达产品，利用随机森林和LightGBM算法实现了对冰雹等强对流天气的分类预警和预报。

冰雹作为一种灾害性天气，具有很强的局地性，各地冰雹云的特点和活动规律不同，其识别参量也不尽相同[10]。内蒙古东西狭长，气候环境差异

收稿日期：2024-01-25； **修订日期**：2024-04-25

基金项目：国家自然科学基金重点项目（42030604）；内蒙古自然科学基金面上项目（2024MS0424）；中国气象局创新发展专项项目（CXFZ2022J033）；内蒙古自治区气象局科技创新项目（nmqxkjcx202470）资助

作者简介：辛悦（1995-），女，硕士，工程师，主要从事大气物理学与人工影响天气研究. E-mail: xinqzh163@163.com

通讯作者：苏立娟（1976-），女，硕士，正研级高工，主要从事大气物理学与人工影响天气研究. E-mail: 13847146506@163.com

较大，地形复杂多样，大范围冰雹、雷暴、大风等强对流天气时有发生，对于内蒙古而言冰雹灾害是仅次于干旱的第二大灾害[11]。因此，本文基于内蒙古人工观测冰雹记录，运用统计方法对1959—2021年内蒙古冰雹灾害事件的变化规律和时空分布特征进行研究，通过多种机器学习方法进行建模和冰雹识别，对比分析不同机器学习方法的识别效果，以期进一步提高冰雹的预报预警准确率，为内蒙古冰雹预报预警和人工防雹作业提供参考。

1 资料与方法

1.1 研究区概况

内蒙古(37°24′~53°23′N，97°12′~126°04′E)地处我国北部边疆，总面积达1.18×10⁶ km²。下辖9个地级市和3个盟，共12个盟市，其中阿拉善盟、乌海市、鄂尔多斯市、巴彦淖尔市和包头市属于内蒙古西部地区，呼和浩特市、乌兰察布市和锡林郭勒盟属于内蒙古中部地区，赤峰市、通辽市、兴安盟和呼伦贝尔市属于内蒙古东部地区[12]。内蒙古地貌格局呈现平原、山地、高原的带状分布，以高原为主，高原四周分布着大兴安岭、阴山山脉。内蒙古所处的地理位置和地形特点，常年受中纬度西风带气流影响，主要以温带大陆性季风气候为主，时常引起冷暖气流交汇，是冰雹等强对流天气频发区域。

1.2 资料来源

选用1959—2021年地面气象观测站记录的特殊天气实况观测资料、2009—2018年巴彦淖尔市地面防雹作业点的降雹记录及对应时刻多普勒雷达资料、1959—2021年地面观测的小时雨量资料，用于筛选冰雹及非冰雹个例。同时，使用ERA5再分析资料，用于计算各种天气条件下大气环境场的热力、动力、水汽参数，作为机器学习时的预报因子。

1.3 机器学习标签数据集构建

冰雹天气正样本标签主要来自于地面气象观测站记录的冰雹过程以及经过雷达回波产品校对的防雹作业点的降雹记录。由于强对流天气经常相伴发生，在筛选冰雹个例时，根据灾害影响程度对强天气进行侧重分类，对于冰雹和雷暴均出现的情况，记为冰雹过程，对于冰雹和大风同时出现的情况，一般记为冰雹过程。由于防雹作业点的降雹记录只记载了各乡镇冰雹发生时间，而无具体降雹点位置，对照冰雹发生时间的多普勒雷达相关产品，逐个确定降雹点具体经纬度，认为雷达组合反射率因子最强的地点记为该行政区域内的降雹点。共计收集到正样本10066个。

负样本标签主要由雷暴、大风、短时强降水、一般性降水4种天气类型个例组成。其中，雷暴和大风观测来自于地面气象观测站记录的特殊天气实况观测资料，短时强降水为地面站点观测到小时雨强大于20 mm·h^{-1}的记为1次短时强降水过程，一般性降水则为根据增雨作业记录筛选的典型降水过程个例。共计收集到负样本7468个。

1.4 机器学习方法

为对比不同机器学习算法对冰雹的识别能力，采用前人研究中常用且实际预测效果较好的随机森林、LightGBM、*K*近邻和决策树4种机器学习算法进行建模试验[8-9,13-14]。本文将冰雹预报预警问题归结为二分类问题，总体的训练思路如下：根据机器学习训练集中的冰雹天气标签及负样本标签数据，利用不同机器学习算法建立特征工程，经过反复的训练调参和交叉验证，确定各模型的超参数组合，不断优化模型算法，采用命中率、空报率、漏报率和TS评分在测试集上进行检验，判断各模型的识别效果。

2 结果与分析

2.1 冰雹特征统计

2.1.1 冰雹的时间分布特征 根据之前研究对冰雹出现次数的定义[5]，规定当某测站在某日观测到1次或1次以上冰雹天气现象时，不论其时间长短作为一条数据记录为1个冰雹站日数，该测站在该日记录为1个雹日，冰雹发生时间以最早记录时间为准。内蒙古每年平均74站出现冰雹，共165站日数。1984年记录到冰雹过程的站数最多，共有101站观测到冰雹过程，1994年仅有36站记录到冰雹过程，为1959—2021年中最少。冰雹站日数与冰雹站数的峰谷值年份并不完全对应，冰雹站日数最多的年份为1976年(324站日数)，冰雹站日数最少的年份为2007年(46站日数)。

1959—2021年内蒙古119个站的冰雹站数和冰雹站日数的年际变化如图1所示。内蒙古冰雹发生整体呈现下降趋势，20世纪60年代到90年代，冰雹

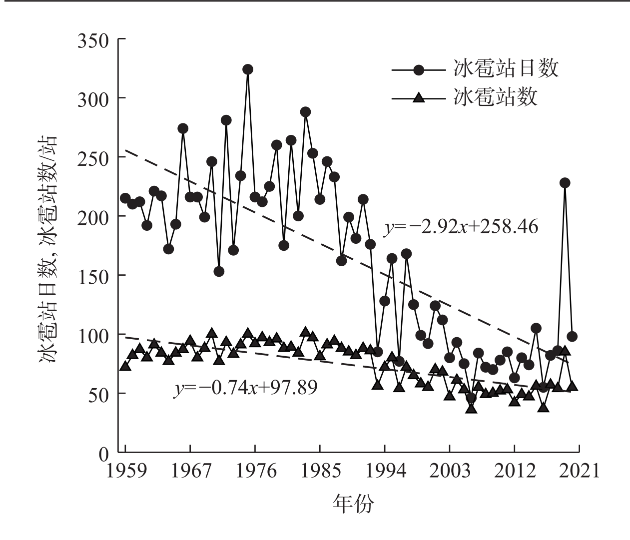


图1　1959—2021年内蒙古冰雹站日数及冰雹站数的年际变化

Fig. 1　Interannual variations of hail station days and hail station numbers in Inner Mongolia from 1959 to 2021

高发，冰雹站日数平均可达222，冰雹站数平均为86站，21世纪以后，冰雹发生显著减少，平均冰雹站日数为86，平均冰雹站数为53站，但是2020年冰雹发生次数突然急剧增加，全区83个站出现冰雹，冰雹站日数达到228。根据变化趋势线显示，冰雹站日数年均减少2.9，冰雹站数年均减少0.7站，与我国冰雹频次整体呈现下降的趋势相同[4-5]。在全球变暖的大背景下，内蒙古平均气温存在明显的升温趋势，线性变化倾向率为0.38 ℃·$(10a)^{-1}$，而冰雹站日数和冰雹站数均呈现下降趋势，表明在地形不变的条件下，气候变暖可能造成大气环流和下垫面性质发生变化，从而改变区域水汽条件、层结稳定度和抬升力大小，导致强对流事件发生变化[15-16]。

1959—2021年冰雹的月分布特征如表1所示，全年3—12月均有冰雹发生，1—2月未观测到冰雹过程。冰雹主要集中在5—9月，占全年的91.79%，其中冰雹站日数排在前3位的分别为6月(2645站日数)、7月(2134站日数)、8月(1697站日数)，分别占全年冰雹过程的25.45%、20.54%和16.33%。从季节变化特征看，夏季(6—8月)冰雹站日数占到全年50%以上，达62.32%，其次是秋季(9—11月)，占19.41%，最少的是冬季(12—翌年2月)，63 a仅发生过2次冰雹过程。

1959—2021年冰雹站日数的日分布情况如图2所示，冰雹日分布呈现单峰结构，00:00—08:00冰雹发生较少，10:00以后冰雹发生频率迅速增加，于15:00达到峰值(16.3%)，20:00以后又迅速减少。冰雹主要发生在12:00—18:00，占全天的80%以上，午后由于太阳辐射加强、气温升高、气压下降、近地层大气容易产生热力不稳定，出现中小尺度的强对流天气，发生冰雹天气。00:00—08:00以及21:00—24:00是冰雹发生较少的时段，夜间和清晨大气层结稳定，不易发生对流。内蒙古冰雹日变化属于午后多雹型，与邸强等[17]、哈青辰[18]对呼和浩特市、赤峰市的冰雹日变化特征的研究结果较为一致。

2.1.2　冰雹的空间分布特征　在119个地面气象观测站中，单站冰雹日数排在前3位的分别为乌兰察布市察哈尔右翼前旗(254 d)、呼和浩特市武川县(245 d)、锡林郭勒盟太仆寺旗(212 d)。统计分析119站63 a平均的年冰雹日数，年均冰雹日数在3 d以上的共有5站，均位于内蒙古中部地区，达到2 d以上的有23站，年均1 d及以上的有76站，而有冰雹记录但年均冰雹日数不到1 d的有43站，占全部降雹测站的36.13%。多雹区主要位于阴山山脉和大兴安岭一带，沿山脉伸展分布。草原地区和荒漠地区的年均冰雹日数则较低，其中阿拉善盟8个测站年均冰雹日数仅为0.32 d。我国多雹区主要位于青藏高原、天山、大小兴安岭、祁连山、贺兰山、阴山等海拔较高、地形复杂的山区[5,10]。这主要是由于海拔高处地形起伏大、植被分布不均，下垫面白天升温剧烈，冷空气过山后易形成山谷风环流，在水汽充足的条件下，利于对流的发生发展，易产生冰雹天气[11,19]。

2.2　基于机器学习的冰雹识别方法

2.2.1　数据预处理　在进行模型训练之前，先将构建的机器学习标签数据集内的正负样本随机打乱，

表1　1959—2021年内蒙古冰雹站日数及冰雹站日数频率月变化

Tab. 1　Monthly variations of the hail station days number and frequency in Inner Mongolia from 1959 to 2021

冰雹特征参数	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
冰雹站日数	0	0	30	448	1419	2645	2134	1697	1644	364	9	2
冰雹站日数频率/%	0.00	0.00	0.29	4.31	13.65	25.45	20.54	16.33	15.82	3.50	0.09	0.02

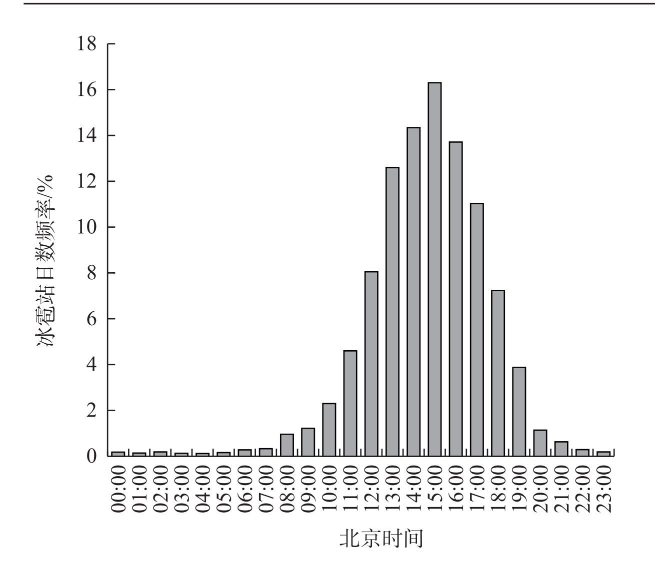


图2 1959—2021年内蒙古冰雹站日数频率日变化

Fig. 2 Daily variations of hail station days frequency in Inner Mongolia from 1959 to 2021

并按照80%和20%的比例划分为训练集和测试集，其中，训练集样本量14027个，冰雹正样本8075个，非冰雹负样本5952个，测试集样本量3507个，冰雹正样本1991个，非冰雹负样本1516个。为避免预报因子的量级差异对模型训练结果造成影响，在模型训练之前对所有特征量进行了标准化处理。

2.2.2 预报因子选择与重要度分析 根据选取的机器学习正负样本标签数据，得到各类正负样本发生的具体时间及对应位置，基于ERA5再分析资料计算各样本环境场的热力、动力、水汽条件及各类强对流天气指数作为预报因子。经过数据筛选和清洗，剔除特征变量不全或存在奇异值的样本，最终形成了基于数值模式分析场资料计算的若干对流指数和诊断量场，共计31类物理量作为机器学习模型的备选预报因子，具体参数类型如表2所示。

决策树、随机森林和LightGBM通过计算信息熵从而建立特征工程，寻找与预报结果关系最为紧密的预报因子对样本进行分类，并可以输出各预报因子对建模的贡献度[20]，贡献度排名前8的预报因子如图3所示。在所有模型中，总指数和冰雹发生时间对模型的贡献度最大，均位列前二，其次对冰雹分类较为重要的因子主要为柱水汽量和柱总水量等水汽条件，850~500 hPa风切变和10 m瞬时最大风速等动力因子，以及850~500 hPa温度差、−20~0 ℃厚度、−20 ℃层高度、对流有效位能等不稳定条件。由此可见，机器学习算法筛选的因子的物理意义较为明确，和主观预报经验基本相符，因此利用机器学习方法建立冰雹判别模型可信度较高，可用于业务应用。

机器学习自变量因子的选择对模型的判别结果有一定的影响，其影响取决于该因子对模型的贡献度。根据以上各预报因子的贡献度排序，选取贡献度较大的*K*指数、总指数、500 hPa假相当位温、850~500 hPa假相当位温差、850~500 hPa温度差、700 hPa露点温度差、−20~0 ℃厚度、−20 ℃层高度、对流有效位能、柱总水量、柱水汽量等11个预报因子加上冰雹发生的时间地点等基础信息作为模型训练的自变量，开展模型训练。根据筛选模型的自变量因子时所做的测试，增加或减少某一自变量得到的机器学习训练结果均低于目前的识别效果。

2.2.3 机器学习训练结果 分别利用随机森林、LightGBM、*K*近邻和决策树4种机器学习算法在训练集使用默认参数进行训练，并对训练得到的模型在测试集上进行识别，识别结果如表3所示。

4种机器学习算法中，随机森林的识别效果最好，1991个冰雹样本，预报正确1854个，漏报了137

表2 机器学习训练的主要预报因子类型和要素

Tab. 2 Types and elements of the main predictive factors in machine learning training

预报因子类型	预报物理量
指数类	*K*指数、修正*K*指数、总指数、强天气威胁指数、沙氏指数
水汽因子	柱总水量、柱水汽量、850 hPa相对湿度、700 hPa相对湿度、850 hPa露点温度差、700 hPa露点温度差、瞬时最大湿通量
动力因子	850~500 hPa风切变、850~300 hPa风切变、10 m瞬时最大风速
高度层因子	0 ℃层高度、−20 ℃层高度、−20~0 ℃厚度
热力因子	850~500 hPa温度差、850 hPa假相当位温、700 hPa假相当位温、500 hPa假相当位温、850~500 hPa假相当位温差、700~500 hPa假相当位温差、对流有效位能
其他信息	冰雹发生的年份、月份、日期、当地时、经度、纬度

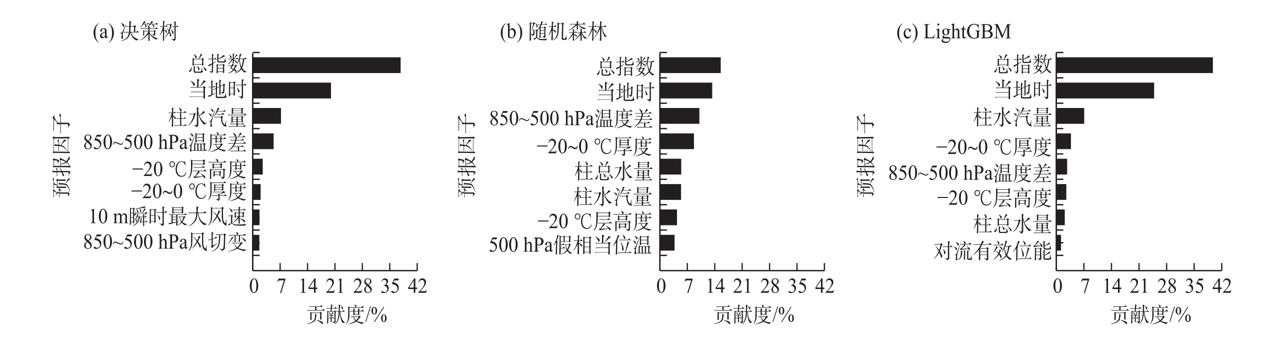


图3　各机器学习算法贡献度排名前8的预报因子重要性排名

Fig. 3　Importance ranking of the first eight variables of each machine learning algorithm

表3　默认参数训练的机器学习算法在测试集上的识别效果

Tab. 3　Recognition performance of machine learning algorithms trained with default parameters on the test set

模型名称	命中数/个	漏报数/个	空报数/个	命中率/%	空报率/%	漏报率/%	TS评分
随机森林	1854	137	158	93.11	7.82	6.88	0.8627
LightGBM	1820	171	168	91.55	8.59	8.45	0.8429
*K*近邻	1777	214	154	89.25	7.97	10.74	0.8284
决策树	1759	232	217	88.34	10.98	11.65	0.7966

个，命中率为93.11%，空报率为7.82%，漏报率为6.88%，TS评分为0.8627，在所有模型中其命中率和TS评分均最高，空报率和漏报率均最低。LightGBM模型冰雹命中率也在90%以上，*K*近邻算法次之，决策树排名最后，但是其命中率也达到了88.34%，只是空报率和漏报率比其他模型高，TS评分也达到了0.7966，整体识别效果优于其他的强对流天气分类落区预报产品[21]。

2.2.4　*模型调优*　在机器学习模型训练中，超参数调优对模型的训练效果有重要影响，不同的超参数组合可以显著影响模型的性能[22]。机器学习超参数调优是为了找到一组最佳的超参数组合，使模型在特定任务上表现最佳，其对于提高模型性能、防止过拟合、加速收敛等方面都非常重要[23]。在*K*近邻算法中，最重要的超参数为n_neighbors，即分类时选取临近样本的个数，该值过小时，会使得分类结果对异常值更敏感，容易受到噪声干扰导致过拟合，该值较大时，会使分类结果对临近样本的影响更大导致欠拟合。决策树、随机森林和LightGBM这类树模型中，n_estimators、max_depth和min_sample_leaves都是重要的超参数指标。n_estimators代表基本学习器的数量，是随机森林和LightGBM模型中最重要的参数，通常情况下，该值越大，模型性能越好，但是也会导致训练时间的增加和过拟合的问题。max_depth代表决策树的最大深度，该值越大模型的复杂度越高，容易出现过拟合情况。min_sample_leaves代表每个叶子节点至少包含的样本数，该值越小，模型的灵活度越高。本文选取的各超参数取值范围如表4所示，通过对不同超参数组合进行试验，测试模型效果。

利用网格搜索方法，对上述取值范围内的超参数组合进行试验，分别对4种模型进行调优，调优后随机森林参数选择为n_estimators=161，LightGBM参数选择为n_estimators=152，*K*近邻参数选择为n_neighbors=4，决策树参数选择为max_depth=8，min_sample_leaves=3，调优后各模型在测试集上的识别结果如表5所示。调优后各模型的识别效果较调优前都有一定的改善，各模型TS评分提升了0.01~0.04，其中决策树模型训练效果提升幅度最大，其命中率提高了4.78%，TS评分提升了0.0396；但

表4　机器学习模型超参数调节取值范围

Tab. 4　Range of hyperparameter tuning values for machine learning models

超参数	初始值	取值范围
n_neighbors	3	[1, 11]
n_estimators	100	[10, 500]
max_depth	2	[1, 11]
min_sample_leaves	10	[1, 50]

是各模型识别效果排序没有变化,TS评分由高到低依次为随机森林、LightGBM、*K*近邻和决策树。

2.2.5 模型应用效果检验 利用2023年5—8月内蒙古地面人工观测站观测到的冰雹记录,对调优后的模型识别效果进行检验,具体降雹站点及降雹时间和各模型识别情况如表6所示。2023年5—8月共观测到冰雹过程23次,随机森林、LightGBM、*K*近邻和决策树分别报出17次、15次、19次、19次冰雹过程,识别准确率分别为73.91%、65.22%、82.61%和82.61%。针对7月7日林西,8月2日土默特右旗和科尔沁左翼中旗的过程,所有模型均存在漏报的现象,这3次个例均为分散的对流单体导致的冰雹过程,其中土默特右旗的冰雹过程雷达反射率因子最强仅达到30 dBZ,说明模型对较弱的对流单体引起的冰雹事件识别能力略差,后续将添加雷达反射率因子、最大回波顶高、垂直累积液态水等雷达参数,用以加强模型对短时临近过程的冰雹判别能力。

表5 调优后的各机器学习算法在测试集上的识别效果

Tab. 5 Recognition performance of optimized machine learning algorithms on the test set

模型名称	命中数/个	漏报数/个	空报数/个	命中率/%	空报率/%	漏报率/%	TS评分
随机森林	1856	135	155	93.22	7.71	6.78	0.8648
LightGBM	1848	143	158	92.82	7.88	7.18	0.8599
*K*近邻	1855	136	196	93.17	9.56	6.83	0.8481
决策树	1854	137	226	93.12	10.87	6.88	0.8362

表6 2023年5—8月内蒙古冰雹观测记录及机器学习模型识别效果

Tab. 6 Hail observation records and machine learning model recognition performance in Inner Mongolia from May to August of 2023

降雹站点	降雹日期	降雹时间	随机森林	LightGBM	*K*近邻	决策树
正蓝旗	5月11日	16:25—16:28	√	√	√	√
镶黄旗	5月12日	15:56—15:57	√	√	√	√
宁城	5月16日	15:41—15:44	√	√	√	√
正蓝旗	6月1日	12:16—12:18	√	√	√	√
新巴尔虎右旗	6月2日	13:06—13:17	√	√	√	√
扎兰屯	6月2日	12:30	×	×	√	√
胡尔勒	6月14日	19:36—19:37	√	√	√	√
呼和浩特	6月19日	15:41—15:44	√	√	√	√
乌拉特前旗	6月25日	17:16—17:24	√	√	√	√
扎兰屯	6月28日	18:44—18:49	√	√	√	√
林西	7月7日	11:10	×	×	×	×
青龙山	7月7日	10:07	√	√	√	√
科尔沁左翼后旗	7月14日	14:48—14:50	×	×	√	√
东乌珠穆沁旗	7月18日	16:46—16:51	√	√	√	√
吉兰泰	7月20日	12:54	√	√	√	√
赛罕区	7月23日	19:23—19:30	√	×	√	√
呼和浩特	7月24日	15:22—15:55	√	√	√	√
土默特右旗	8月2日	16:33—16:39	×	×	×	×
科尔沁左翼中旗	8月2日	17:06—17:07	×	×	×	×
察哈尔右翼前旗	8月6日	19:20	√	×	×	×
正蓝旗	8月8日	16:22	√	√	√	√
多伦	8月9日	15:34—15:45	×	×	√	√
多伦	8月20日	17:55—18:05	√	√	√	√

注:√表示识别出该次冰雹过程;×表示未识别出该次冰雹过程。

3 讨 论

冰雹天气是内蒙古灾害性天气中较为常见的强对流天气，其来势猛、局地性强、破坏性大，对社会经济和人民财产安全造成了严重威胁。内蒙古地域狭长，地形地貌复杂，气候条件差异较大，冰雹分布也呈现出明显的地域特征，在阴山、大兴安岭等山地地区，冰雹出现频率明显高于其他地区，与汤兴芝等[5]对全国冰雹分布研究中结论相似，多雹区大体沿山系伸展，内蒙古中东部地区降雹频率也居全国前列。

通过长时间序列的分析，发现内蒙古冰雹站日数呈现下降趋势，与新疆、陕西等地[24,11]变化趋势一致，全国大部分地区冰雹出现均呈现下降趋势，这与全球变暖平均气温升高、人工防雹作业能力提升等关系密切[5,25]。内蒙古降雹具有季节性强、雹日高度集中的单峰型月际分布特征，这一结论与虎雅琼等[25]、尉英华等[6]给出的华北地区降雹季节分布特征基本一致，究其原因，主要是北方地区从夏季开始环流形势变得活跃，水汽输送不断加强，在午后太阳辐射增强，地面温度升高，使得边界层大气层结变得很不稳定，利于对流天气发生发展。

然而，冰雹的准确识别和预报预警一直以来都是天气预报面临的最大难题之一，2010—2015年国家级强对流天气主观综合预报产品对6~24 h时效预报雷暴大风和冰雹的TS评分在0.01~0.07[26]。本文利用多种机器学习算法，建立了冰雹识别方法，经过参数调优后，各模型的TS评分均达到0.83以上，命中率达到92%以上，4种机器学习算法在测试集上TS评分由高到低依次为随机森林、LightGBM、*K*近邻和决策树。通过收集并统计大量冰雹数据及对应环境场信息构建的机器学习模型，得到了较为客观科学的冰雹预报结论，有效提升了识别准确率，但是仍存在一定局限性。目前主要利用观测站点人工观测记录的冰雹个例进行冰雹个例识别，受限于人工观测的范围和距离，大量发生在测站范围之外的冰雹过程则无法得到有效记录，未来将采用更多观测数据，如卫星、雷达资料等，弥补监测范围的不足，进一步优化和改进样本数据集，提升模型的预报性能。

4 结 论

（1）1959—2021年，内蒙古冰雹站日数呈减少趋势，同时降雹的站数也呈线性减少趋势，冰雹发生频率和发生范围正在逐步减小。冰雹分布具有明显的地域特征，多雹区主要位于阴山山脉和大兴安岭一带，沿山脉伸展分布。冰雹发生的季节性和日变化特征明显，每年5—9月为冰雹多发时段，冰雹日分布呈现单峰结构，15:00发生冰雹的频率最高占全天的16.3%。

（2）经过机器学习预报因子重要度排序后发现，*K*指数、总指数、500 hPa假相当位温、850~500 hPa假相当位温差、850~500 hPa温度差、700 hPa露点温度差、−20~0 ℃厚度、−20 ℃层高度、对流有效位能、柱总水量、柱水汽量等环境诊断量对冰雹发生发展有重要指示意义，在预报冰雹中起到关键作用。

（3）4种机器学习算法在实际应用中都较好地实现了冰雹识别，训练模型经过超参数调优后，冰雹识别的TS评分均达到0.83以上，命中率达到92%以上，4种机器学习算法在测试集上综合评分由高到低依次为随机森林、LightGBM、*K*近邻和决策树。

参考文献（References）

[1] 姚展予, 屠琦, 安琳, 等. 冰雹形成过程及人工防雹研究综述[J]. 气象学报, 2022, 80(6): 835-863. [Yao Zhanyu, Tu Qi, An Lin, et al. Review of advances in hail formation process and hail suppression and hail suppression research[J]. Acta Meteorologica Sinica, 2022, 80(6): 835-863.]

[2] 陶涛, 张立新, 桑建人, 等. 六盘山区一次非典型冰雹天气过程微物理量特征的分析[J]. 干旱区地理, 2020, 43(2): 299-307. [Tao Tao, Zhang Lixin, Sang Jianren, et al. A case analysis of microphysical characteristics of atypical hail formation over Liupan Mountain, China[J]. Arid Land Geography, 2020, 43(2): 299-307.]

[3] 王昀, 谢向阳, 马禹, 等. 天山北侧成灾雹云移动路径及预警指标的研究[J]. 干旱区地理, 2017, 40(6): 1152-1164. [Wang Yun, Xie Xiangyang, Ma Yu, et al. Moving paths and nowcasting indicators of radar of hail cloud in northern Tianshan Mountains[J]. Arid Land Geography, 2017, 40(6): 1152-1164.]

[4] 张芳华, 高辉. 中国冰雹日数的时空分布特征[J]. 南京气象学院学报, 2008, 31(5): 687-693. [Zhang Fanghua, Gao Hui. Temporal and spatial features of hail days in China[J]. Journal of Nanjing Institute of Meteorology, 2008, 31(5): 687-693.]

[5] 汤兴芝, 黄治勇, 张荣, 等. 2010—2020年全国冰雹灾害事件时空分布特征[J]. 暴雨灾害, 2023, 42(2): 223-231. [Tang Xingzhi, Huang Zhiyong, Zhang Rong, et al. Temporal and spatial distribution characteristics of hail disaster events in China from 2012 to 2020[J]. Torrential Rain and Disasters, 2023, 42(2): 223-231.]

[6] 尉英华, 花家嘉, 王莹, 等. 近11年天津冰雹统计特征及对流参数指标分析[J]. 气象, 2023, 49(2): 213-223. [Wei Yinghua, Hua Jiajia, Wang Ying, et al. Statistical characteristics and convection indicators of hailstorm over Tianjin in recent 11 years[J]. Meteorological Monthly, 2023, 49(2): 213-223.]

[7] 钟敏, 郭英莲, 陈璇, 等. 基于客观分型的冰雹概率预报方法研究[J]. 高原气象, 2022, 41(4): 934-944. [Zhong Min, Guo Yinglian, Chen Xuan, et al. Study on hail probability forecast method based on objective classification[J]. Plateau Meteorology, 2022, 41(4): 934-944.]

[8] 刘新伟, 黄武斌, 蒋盈沙, 等. 基于LightGBM算法的强对流天气分类识别研究[J]. 高原气象, 2021, 40(4): 909-918. [Liu Xinwei, Huang Wubin, Jiang Yingsha, et al. Study of the classified identification of the strong convective weathers based on the LightGBM algorithm[J]. Plateau Meteorology, 2021, 40(4): 909-918.]

[9] 刘新伟, 蒋盈沙, 黄武斌, 等. 基于雷达产品和随机森林算法的冰雹天气分类识别及预报[J]. 高原气象, 2021, 40(4): 898-908. [Liu Xinwei, Jiang Yingsha, Huang Wubin, et al. Classified identification and nowcast of hail weather based on radar products and random forest algorithm[J]. Plateau Meteorology, 2021, 40(4): 898-908.]

[10] 汤兴芝, 黄兴友. 冰雹云的多普勒天气雷达识别参量及其预警作用[J]. 暴雨灾害, 2009, 28(3): 261-265. [Tang Xingzhi, Huang Xingyou. Doppler radar identification parameters and their effect on early warning of hail clouds[J]. Torrential Rain and Disasters, 2009, 28(3): 261-265.]

[11] 韩经纬, 王海梅, 乌兰, 等. 内蒙古雷暴、冰雹灾害的评估分析与防御对策研究[J]. 干旱区资源与环境, 2009, 27(7): 31-38. [Han Jingwei, Wang Haimei, Wu Lan, et al. The analysis and assessment on thunderstorm and hail disasters and the countermeasures in Inner Mongolia[J]. Journal of Arid Land Resources and Environment, 2009, 27(7): 31-38.]

[12] 顾润源, 孙永刚, 韩经纬, 等. 内蒙古自治区天气预报手册[M]. 北京: 气象出版社, 2012: 277-281. [Gu Ruiyuan, Sun Yonggang, Han Jingwei, et al. Weather forecast manual of Inner Mongolia[M]. Beijing: China Meterological Press, 2012: 277-281.]

[13] 李文娟, 赵放, 郦敏杰, 等. 基于数值预报和随机森林算法的强对流天气分类预报技术[J]. 气象, 2018, 44(12): 1555-1564. [Li Wenjuan, Zhao Fang, Li Minjie, et al. Forecasting and classification of severe convective weather based on numerical forecast and random forest algorithm[J]. Meteorological Monthly, 2018, 44(12): 1555-1564.]

[14] 周康辉. 基于深度卷积神经网络的强对流天气预报方法研究[D]. 北京: 中国气象科学研究院, 2021. [Zhou Kanghui. Convective weather forecasting with convolutional neural networks[D]. Beijing: University of Chinese Academy of Sciences, 2021.]

[15] 王芝兰, 陈录元, 尚可政, 等. 青海强对流天气时空特征及其对气候变暖的响应[J]. 干旱气象, 2011, 29(4): 439-445, 454. [Wang Zhilan, Chen Luyuan, Shang Kezheng, et al. Characteristics of temporal and spatial distribution of severe convective weather and its response to climate warming in Qinghai Province[J]. Arid Meteorology, 2011, 29(4): 439-445, 454.]

[16] 加勇次成, 次旺顿珠, 措姆. 气候变化背景下羌塘自然保护区冰雹日数时空变化特征[J]. 中国农学通报, 2019, 35(18): 103-109. [Jiayongcicheng, Ciwangdunzhu, Cuomu. Hail days in Qiangtang National Nature Reserve of Tibet: Spatial and temporal variation characteristics under the background of climate change[J]. Chinese Agricultural Science Bulletin, 2019, 35(18): 103-109.]

[17] 邸强, 穆超, 孔文甲. 呼和浩特地区冰雹时空特征分析[J]. 内蒙古科技与经济, 2023, 4: 92-93, 157. [Di Qiang, Mu Chao, Kong Wenjia. Analysis of spatiotemporal characteristics of hail in Hohhot area[J]. Inner Mongolia Science Technology and Economy, 2023, 4: 92-93, 157.]

[18] 哈青辰. 近32年赤峰地区冰雹的分布特征及防御[J]. 农业灾害研究, 2021, 11(7): 83-84. [Ha Qingchen. Distribution characteristics and defense of hail in Chifeng area in recent 32 years[J]. Journal of Agricultural Catastropholgy, 2021, 11(7): 83-84.]

[19] 罗喜平, 廖波, 张小娟, 等. 1961—2020年贵州冰雹气候特征[J]. 干旱气象, 2022, 40(6): 1024-1032. [Luo Xiping, Liao Bo, Zhang Xiaojuan, et al. Climatic characteristics of hail in Guizhou from 1961 to 2020[J]. Journal of Arid Meteorology, 2022, 40(6): 1024-1032.]

[20] Friedman J H. Greedy function approximation: A gradient boosting machine[J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[21] 张勇, 刘慧, 郑颖菲, 等. 人工智能模型的分类临近预报产品效果检验与分析[J]. 沙漠与绿洲气象, 2023, 17(1): 115-121. [Zhang Yong, Liu Hui, Zheng Yingfei, et al. Effect validation and analysis of classified products outputted by artificial intelligent nowcasting model[J]. Desert and Oasis Meteorology, 2023, 17(1): 115-121.]

[22] 刘瑞亮, 贾科利, 李小雨, 等. 组合光学和微波遥感的耕地土壤含盐量反演[J]. 干旱区地理, 2024, 47(3): 433-444. [Liu Ruiliang, Jia Keli, Li Xiaoyu, et al. Inversion of soil salt content by combining optical and microwave remote sensing in cultivated land[J]. Arid Land Geography, 2024, 47(3): 433-444.]

[23] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[24] 朱思华, 罗继, 曲良璐. 新疆阿克苏地区冰雹时空分布及雷达回波特征[J]. 沙漠与绿洲气象, 2021, 15(2): 81-88. [Zhu Sihua, Luo Ji, Qu Lianglu. The spatial-temporal distribution and radar echo signatures of hail in Aksu, Xinjiang[J]. Desert and Oasis Meteorology, 2021, 15(2): 81-88.]

[25] 虎雅琼, 边宇轩, 黄梦宇, 等. 基于灾情信息的1981—2017年北

京地区降雹特征[J]. 应用气象学报, 2019, 30(6): 710−721. [Hu Yaqiong, Bian Yuxuan, Huang Mengyu, et al. Characteristics of hailstone distribution based on disaster in Beijing from 1981 to 2017[J]. Journal of Meteorological Science, 2019, 30(6): 710−721.]

[26] 唐文苑, 周庆亮, 刘鑫华, 等. 国家级强对流天气分类预报检验分析[J]. 气象, 2017, 43(1): 67−76. [Tang Wenyuan, Zhou Qingliang, Liu Xinhua, et al. Anlyisis on verification of national severe convective weather categorical forecasts[J]. Meteorological Monthly, 2017, 43(1): 67−76.]

Hail characteristics and hail recognition method based on machine learning in Inner Mongolia

XIN Yue[1], SU Lijuan[1], ZHENG Xucheng[1], LI Hui[1], YI Nana[1], JIN Yuchen[2]

(1. Inner Mongolia Weather Modification Center, Hohhot 010051, Inner Mongolia, China; 2. Inner Mongolia Meteorological Science Institute, Hohhot 010051, Inner Mongolia, China)

Abstract: Based on the manual observation of hail records in Inner Mongolia, China, from 1959 to 2021, the spatial and temporal characteristics of hail distribution are analyzed, and a hail recognition method is constructed based on machine learning algorithms. The results are as follows: (1) Regarding temporal distribution, the number of hail days and affected stations in Inner Mongolia shows a decreasing trend. In terms of spatial distribution, hail events are predominantly concentrated in the Yinshan Mountains and the Greater Hinggan Mountains, with hail-prone areas extending along these mountain ranges. (2) Hail exhibits distinct seasonal and diurnal characteristics. The peak hail months in Inner Mongolia are from May to September, accounting for 91.79% of the annual hail days. The most frequent period for hail occurrences is between 12:00 BST and 19:00 BST. (3) Four machine learning algorithms (random forest, LightGBM, *K*-proximity, and decision tree) are used to model and evaluate hail events in Inner Mongolia through data preprocessing, predictor selection, model training, and tuning. Verification results indicate that machine learning methods effectively identify hail events, with the threat score of each model exceeding 0.83 and hit rates surpassing 92%. Among these, the random forest algorithm demonstrates the best recognition performance on the test set. These findings provide useful references for hail forecasting and artificial hail prevention in Inner Mongolia.

Key words: the number of hail station days; temporal and spatial characteristics; machine learning; hail identification